Context-Paraphrase Enhanced Commonsense Question Answering

Anonymous ACL submission

Abstract

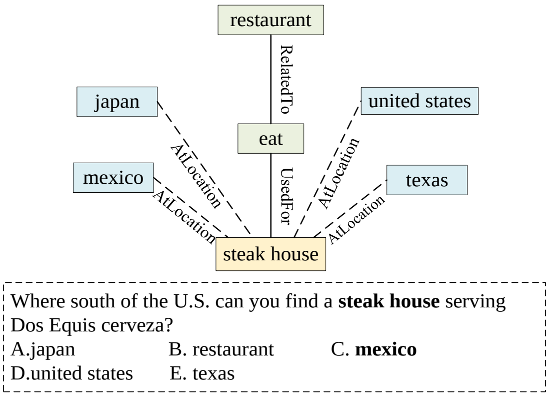
Commonsense question answering (CQA) generally means that the machine use the mastered commonsense to answer questions without relevant background material, which is a challenging task in natural language processing. Many prior methods mainly retrieve question related evidences from the structured knowledge base as the background material of the question, while the extracted evidence is generally described through the entities and the relationship between the entities, making it difficult for the machine to understand the meaning of the evidence completely. In this paper, we integrate the paraphrase in WordNet and Wiktionary into the evidence extraction process and machine reading comprehension (MRC) model, and propose a context-paraphrase enhanced commonsense question answering method. Specifically, the context-paraphrase obtained by WordNet and Wiktionary is first incorporated into the construction process of the heterogeneous graph, and the question related triple is extracted based on the heterogeneous graph, the triple is converted to triple-text based on a relational template. Then, the triple-text is used as the context of the question to establish an association graph containing the relationship between the context entities and the paraphrases. We further integrate the association graph into the MRC model to better guide the model to answer. Experimental results on CommonsenseQA and OpenBookQA show that context-paraphrase is effective in improving the answer accuracy of the MRC model.

1 Introduction

Over the past few years, with the advent of large-scale pre-trained language models (PTLMs)(Devlin et al., 2019; Liu et al., 2019; Lan et al., 2020), MRC tasks have remarkably progressed, surpassing human levels on multiple MRC tasks. However, the PTLMs still have a substantial gap with humans in MRC tasks that require commonsense knowledge despite achieving good results in some tasks. Humans can use their commonsense knowledge in temporal, science, and society to help them understand the meaning of natural language according to practical situations. For example, if you ask: "Where would you expect to find a pizzeria while shopping?", then we know that a "pizzeria" can make "pizza", and "pizza" is a food. Therefore, inferring that you can "find a pizzeria" in a "food court" is easy. This simple reasoning ability may seem easy to human beings but is beyond the current capacity of natural language understanding systems.

Commonsense is the common daily consensus of most people on the same thing, which is the basis of daily human communication and cooperation. Commonsense can be categorized according to types, including social, temporal, physical commonsense, and so on. Several CQA datasets have recently been built on the basis of different types of commonsense. For example, SocialQA(Sap et al., 2019), MCTA-CO(Zhou et al., 2019), PIQA(Bisk et al., 2020), and CommonsenseQA(Talmor et al., 2019) are respectively proposed for social commonsense, temporal commonsense, physical commonsense, and general commonsense.

PTLMs also capture general knowledge about the world. PGFull(Wang et al., 2020) proposed a knowledge path generator to generate structured evidence according to the question dynamically. The generator takes PTLMs as the backbone using a large amount of unstructured knowledge stored in the language model to supplement the incompleteness of the structured knowledge base. However, the knowledge representation in PTLMs remains unclear, and even the knowledge of PTLMs for a particular question may be noise, affecting the response of the machine. Abundant commonsense knowledge is stored in knowledge bases (KBs), and machine can use these KBs to make sound judgments. At the same time, the KBs can also provide displayed and explanatory evidence. Therefore, most CQA methods introduce KBs when solving CQA tasks to improve the commonsense reasoning capability of machines. KagNet(Lin et al., 2019) retrieved the relationship path between the question and choice entities from ConceptNet(Speer et al., 2017) and modeled the relationship between the entity nodes through Graph Attention Networks (GAT)(Veličković et al., 2018) and LSTM. MHGRN(Feng et al., 2020) unified the reasoning methods based on path and Graph neural Networks to achieve improved interpretability and scalability. QAGNN(Yasunaga et al., 2021) used PTLMs to compute the relevance of KB nodes conditioned on the given CQA context, then joint reasoning over the CQA context and KB. Lv et al. (2020) extracted evidences from ConceptNet and Wikipedia, constructed graphs for both sources according to the relationship between evidences, and proposed a graph-based reasoning method to predict the answer.



Where south of the U.S. can you find a **steak house** serving Dos Equis cerveza?
A.japan B. restaurant C. **mexico**
D.united states E. texas

Figure 1: An example from the CommonsenseQA dataset,which requires ConceptNet to make the correct prediction.

Most CQA methods currently attempt to integrate the KBs into the MRC model to improve its reasoning capability. Generally, the KBs (i.e. ConceptNet) store knowledge in the form of triples and lack description of entity paraphrase. An example is shown in Figure 1, in which the relationships between the question and choice entities can be obtained through ConceptNet. But based on these relationships alone, it is difficult for the MRC model to get the correct answer. For humans, we can know that "steak house" means "a restaurant that specializes in serving steak," and "mexico" means "a country in southern North America." Combined with these common senses, the answer C can be easily obtained. However, the MRC model lacks an understanding of the interpretation of these entities, making it difficult to get the correct answer.

The dictionary stores the paraphrases of words and phrases, which are helpful for MRC model answering. Therefore, we propose the context-paraphrase enhanced commonsense question answering (CPE) method. First, the question and choice entities are identified, and the paraphrases of these entities in the dictionary are retrieved. At the same time, the question related triple-texts in ConceptNet are also obtained on the basis of question and choice entities. And based on this, we build a heterogeneous graph with different types of nodes and edges, which is named as Heterogeneous Triple-Paraphrase(HTP) graph. Next, we propose a HTP-based evidence extraction method. We extract the most relevant triple-text to the question on the HTP graph through GAT, and take it as the context of commonsense question. Finally, the paraphrase association graph is established in accordance with the relationship between the context entity and its corresponding paraphrase, and the association graph is incorporated into the answering process of the MRC model.

The main contributions of this paper are as follows:

(1) We incorporate the paraphrase of the entity into the construction process of the HTP graph and propose a HTP-based evidence extraction method.

(2) We construct a paraphrase association graph, and integrate it into the MRC model, to weaken the influence of the entity paraphrase on the non-associated entity.

(3) We also evaluate our method on CommonsenseQA and OpenBookQA, and prove the effectiveness of the proposed method through a series of ablation experiments.

2 Methods

2.1 Problem formulation

For a CommonsenseQA(Talmor et al., 2019) task, given a question and choices $A = \{a_1, a_2, \ldots, a_p\}$, the MRC model needs to choose the correct answer from A without background material. Therefore, the external KBs can provide some useful information for commonsense question. We use ConceptNet(Speer et al., 2017), WordNet(Miller, 1995), and Wiktionary[1] as external KBs.

We propose a HTP-based evidence extraction method to retrieve the most relevant triple-text from

[1] https://www.wiktionary.org/

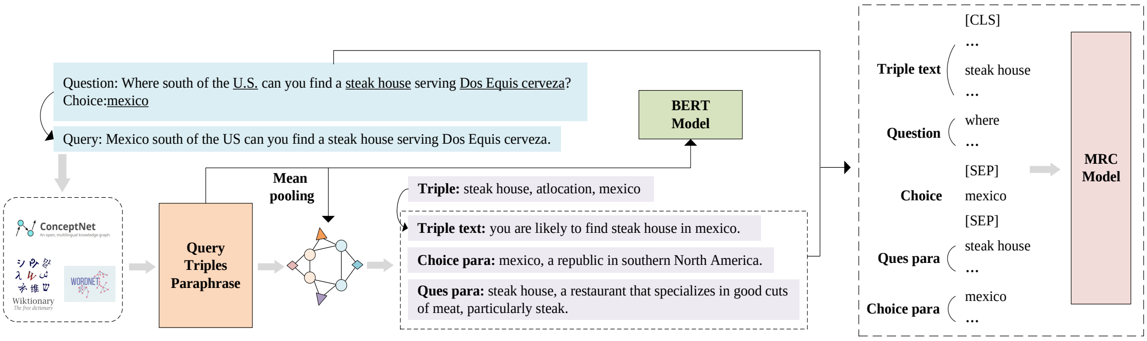

Question: Where south of the U.S. can you find a steak house serving Dos Equis cerveza?
Choice:mexico
Query: Mexico south of the US can you find a steak house serving Dos Equis cerveza.
ConceptNet
Wiktionary
WORDNET
Query Triples Paraphrase
Mean pooling
BERT Model
Triple: steak house, atlocation, mexico
Triple text: you are likely to find steak house in mexico.
Choice para: mexico, a republic in southern North America.
Ques para: steak house, a restaurant that specializes in good cuts of meat, particularly steak.
[CLS]
Triple text
steak house
Question
where
[SEP]
Choice
mexico
[SEP]
Ques para
steak house
Choice para
mexico
MRC Model

Figure 2: HTP-based evidence extraction, the circle represents the entity node, and the same color represents that these entities appear in the same triple-text; the triangle represents the paraphrase node, and the diamond represents the triple-text node.

Table 1: Example of a template conversion from triple to triple-text.

Triple	Template	Triple-text
revolving door-**AtLocation** -bank	You are likely to find [A] in [B].	You are likely to find revolving door in bank.
playing guitar-**Causes** -hear sounds	Sometimes [A] causes [B].	Sometimes playing guitar causes hear sounds.
communicating-**HasSubevent** -learning	Something you might do while [A] is [B].	Something you might do while communicating is learning.

ConceptNet based on the question entity, choice entity, and entity paraphrase. The triple-text is taken as the context for commonsense question. Simultaneously, we construct a paraphrase association graph in accordance with the relationship between the context entity and its paraphrase, and incorporate it into the MRC model, weakening the effect of the entity paraphrase on the non-associated entity. Next, each module is comprehensively described.

2.2 HTP-based evidence extraction

ConceptNet is a large-scale knowledge base of commonsense comprising relationship-based knowledge in the form of triples, with millions of nodes and relationships. Following a previous study Lin et al. (2019). As shown in Figure 2, the question and choice entities are first identified, and the question related triples in ConceptNet are then retrieved on the basis of the question and choice entities [2]. Instead of picking all triples, the top x triples with the highest triple weights are selected in this paper. As shown in Table 1, a relationship transformation template is designed on the basis of the relationship definition in ConceptNet which can convert triples to triple-texts to describe triples effectively.

Understanding the meaning of context based solely on the relationship between entities in context is difficult for the model. Therefore, the relevant paraphrases of these entities in WordNet and Wiktionary are retrieved: if the entity is a single word, then the first paraphrase in WordNet is selected as the relevant description of the entity; if the entity is a phrase, then the first paraphrase in Wiktionary is selected as the relevant description of the entity, and if the phrase is not in Wiktionary, we further search WordNet for word paraphrase of nouns and verbs in that phrase. Then, we further construct a HTP graph based on question, choice, triple-texts and paraphrases. Inspired by the work of DFGN(Qiu et al., 2019), we extract the triple-text most relevant to the question based on HTP graph. The triple-text is taken as the context for the question.

The construction of HTP graph and the extractions of the triple-text are further described below.

Constructing HTP Graph: The interrogative in the question is replaced with a choice and used as a query. The HTP graph is incorporated by the relationship among query, triple-texts, and paraphrases, which include query entity node, triple entity node, paraphrase node and triple-text node.

[2] In CommonsenseQA, the question and option entities provided by the dataset and the named entities recognized by the spacy are used, and we extract one-hop triples in ConceptNet; in OpenBookQA, the nouns, verbs, and named entities identified by the spacy are extracted as entities, and we extract triples within two-hops in ConceptNet.

The following rules are used in the construction of the graph:

(1) a triple entity node and a query entity node are connected, if they are the same;

(2) the two entity nodes are connected, if they come from the same triple-text;

(3) the two entity nodes are connected, if they are the same and come from different triple-texts;

(4) an entity node and a paraphrase node are connected, if the paraphrase node is a definition of this entity;

(5) a triple node and an entity node are connected, if the entity appears in the triple-text at least one time.

Encoding Entity: Given a graph structure, the next step is to obtain an initial representation of each node in the graph, with a pretrained BERT as the node encoder. First, the query, triple-texts and entity paraphrases are fed into the BERT model as an input, and the representation of each token in the input $C = \{c_1, c_2, ..., c_l\} \in \mathbb{R}^{l \times d}$ is obtained, where l is the length of input, and d is the size of BERT hidden states. For each node in the graph, $e = \{c_i, c_{i+1}, \dots, c_k\}$ is a certain segment of the input, and the node representation is obtained through the average pooling method.

$$v = AvergePooling(e) \qquad (1)$$

Graph reasoning: After obtaining the initial representation of each node in the graph, we apply GAT to inference on the graph. Specifically, reasoning starts with an entity in query, and other entities that are connected to that entity in the graph are emphasized. The character representation of the entity is updated by calculating the attention score between them. Assuming that the neighboring node is N_i for any node i, the attention weight of i is then calculated as:

$$e_{ij} = a^{\mathrm{T}}[Wv_i||Wv_j], j \in N_i \qquad (2)$$

$$\alpha_{i,j} = \frac{\exp\left(LeakyReLu\left(e_{ij}\right)\right)}{\sum_{k \in N_n} \exp\left(LeakyReLu\left(e_{ik}\right)\right)} \qquad (3)$$

where $W \in \mathbb{R}^{F' \times F}$, $a \in \mathbb{R}^{2\ F}$, α_{ij} is the attention weight of the node i to its neighbor entities.Finally, the final character representation of the node i is obtained in accordance with α_{ij}.

$$v_i' = \sum_{j \in N_i} \alpha_{ij} W v_j \qquad (4)$$

Through reasoning on the HTP graph, we can better get the embedding $T = \{T_1, T_2, \dots, T_y\}$ of triple-texts, and the correlation score s of each triple-text regarding query is obtained through a linear layer. Finally, the highest scoring triple-text serves as the context for the question.

$$s = linear(T) \qquad (5)$$

2.3 MRC-based Answer Prediction

Given a question, the triple and entity paraphrase related to the question in ConceptNet, WordNet and Wiktionary are retrieved, and the triple is converted to triple-text based on the defined relationship template, and then the triple-text is then used as the context for commonsense question. Context, question, choice, question entity paraphrase and choice entity paraphrase are inputted into a pretrained RoBERTa as shown in Figure 3. Simultaneously, the paraphrase association graph G is further established in accordance with the relationship between entity paraphrases to weaken the influence of entity paraphrase on non-associated entities. $G_{i,j} = 1$ indicates the presence of edges between the two tokens, while $G_{i,j} = 0$ indicates the absence of edges between the two tokens. Specifically, similar to the previous MRC model based on fine-tuned PTLMs, context, question, and choice are visible to each other, and edges are found between their tokens. The question entity paraphrase and the choice entity paraphrase should only be related to themselves and the associated entities. Therefore, the paraphrases are only used to establish edges between themselves and the tokens in associated entities regardless of other tokens.

G is incorporated into the RoBERTa model, preventing changing the meaning of the other non-associated entities or even entire sentences due to entity paraphrases.

In the RoBERTa model, G is further defined as:

$$g_{i,j} \quad = \quad \begin{cases} -10^4 & G_{i,j} = 1 \\ 0 & G_{i,j} = 0 \end{cases} \qquad (6)$$

We integrate g into the self-attention layer of RoBERTa.

$$s_{i,j} = \frac{Q_t K_t^T}{\sqrt{d}} \qquad (7)$$

$$a_{ij} = softmax(s_{ij} + g_{ij}) \qquad (8)$$

$$h_{t+1} = a_{ij} V_t \qquad (9)$$

The h_t is the hidden state of RoBERTa at t moment. Q_t, K_t and V_t are obtained by linear transformation of h_t through three different fully connected

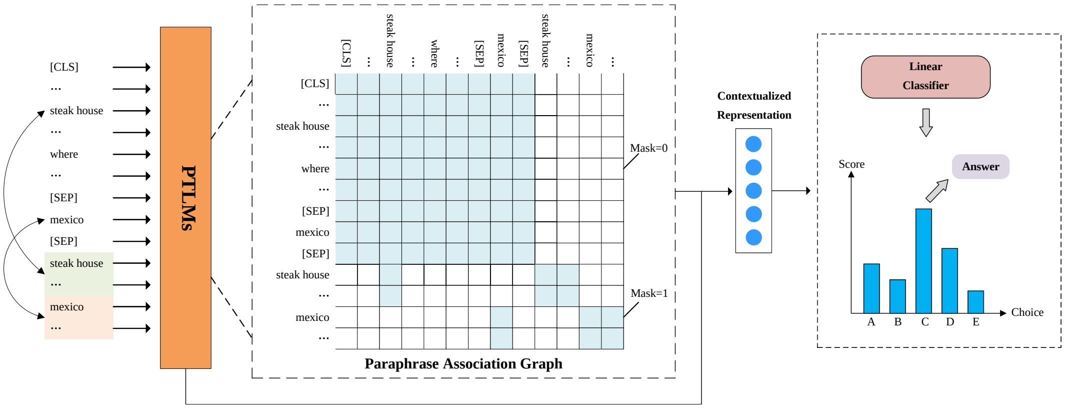


Figure 3: CPE model, wherein the input is triple-text, question, choice, question entity paraphrase, and choice entity paraphrase.

layers. a_{ij} is the attention weight after integrating G.

We use the embedding u of [CLS] as a contextualized representation of the entire input, and a linear classifier is then used to predict the score of the current choice $score(q, u)$.

$$score(q, u) = Linear(u) \tag{10}$$

Finally, the highest-scored choice is chosen as the answer.

3 Experiment

3.1 Datasets

We evaluate our method on two Commonsense Question Answering datasets:

CommonsenseQA(Talmor et al., 2019), a commonsense question answering dataset with multiple-choice questions, requires different types of commonsense knowledge to predict the correct answer. Each question contains one correct choice and four interference choices. CommonsenseQA contains a total of 12,102 questions (train/development/test:9741/1221/1140). The answers to the test set were not officially published. Thus, the work conducted by Lin et al. (2019). is used as a reference, and the train and development sets are divided into in-house dataset (IHdata), where the train set is divided into IHtrain/IHtest: 8500/1241, and the development set is divided into IHdev:1221.

OpenBookQA(Mihaylov et al., 2018) is a question answering dataset based on an open book exam that evaluates human understanding of a topic, and each question must be answered in combination with scientific facts or commonsense knowledge. OpenBookQA contains 5957 multiple-choice questions (train/development/test:4957/500/500). Each question contains one correct choice and four interference choices.

The statistics for the datasets are shown in Table 2.

Table 2: Statistics of CommonsenseQA (CSQA) and OpenBookQA (OBQA).

Datasets	Train	Development	Test
CSQA(Official)	9741	1221	1140
CSQA(IHdata)	8500	1221	1241
OBQA	4957	500	500

3.2 Baselines

We use RoBERTa-large(Liu et al., 2019) to finetune our model on CommonsenseQA and OpenBookQA, and compare with existing RoBERTa-large+KBs methods, including relation network (RN)(Santoro et al., 2017), RGCN(Schlichtkrull et al., 2018), GconAttn(Wang et al., 2019), KagNet(Lin et al., 2019), MHGRN(Feng et al., 2020), and QAGNN(Yasunaga et al., 2021).

For CommonsenseQA, the model is also finetuned on the basis of ALBERT-xxlarge(Lan et al., 2020), and the approach is presented on the leaderboard. Table 4 compares some of the most advanced models (single models) in the leaderboard, including the following:

(1) **RoBERTa + IR**: First, the RoBERTa model is finetuned on the TRACE dataset, and the context information related to the problem is then retrieved through a search engine.

(2) **XLNet + Graph Reasoning**(Lv et al., 2020): Evidences from the two heterogeneous knowledge bases of ConceptNet and Wikipedia is extracted,

and a GNN-based commonsense question answering method is proposed.

(3) **ALBERT + Path Generator**(Wang et al., 2020): The structured knowledge base is supplemented with unstructured knowledge stored in the pre-trained language model.

(4) **RoBERTa + QAGNN**(Yasunaga et al., 2021): The PTLMs is used to compute the relevance of KB nodes conditioned on the given CQA context, then joint reasoning over the CQA context and KB.

(5) **RoBERTa + MHGRN**(Feng et al., 2020): A multihop graph relationship network, which combines the path-based reasoning method with GNN, is proposed.

(6) **ALBERT + KCR**[3]: A triple weight calculation method is designed using ConceptNet as an external knowledge base, and the highest weight triple is chosen as the context of commonsense questions.

(7) **ALBERT + DESCKCR**(Xu et al., 2021): The entity description in Wiktionary is used to provide contextual information for the knowledge graph.

(8) The language models RoBERTa(Liu et al., 2019), ALBERT(Lan et al., 2020), and T5(Raffel et al., 2020) are utilized to fine-tune the train set.

3.3 Experimental Setting

We use the Huggingface(Wolf et al., 2020) implementation for the PTLMs. In the experiment, we set the learning rate to 1e-5, batch size to {4,8}, epoch to 10, and limit the maximum input length to 175. We use Adam(Loshchilov and Hutter, 2019) as the model optimizer.

3.4 Experimental Results

CommonsenseQA: Table 3 shows the results of experiments on IHdata. Our model performs better than baselines. Compared with QAGNN, Our model increased by 2.38% and 1.6% in IHdev and IHtest sets, respectively.

Table 4 shows the accuracy on the official development and test sets of CommonsenseQA. Compared with the baseline models, our proposed CPE achieves the best experimental results on the official development set.

OpenBookQA: Additional experiments on the OpenBookQA data are conducted to further demonstrate the effectiveness of the proposed method. Table 5 shows that the CPE increased 6.2% higher than the RoBERTa model and QAGNN increased 0.42%.

Table 3: Performance comparison on Commonsense QA in-house split.

Methods	IHdev	IHtest
RoBERTa-large(w/o KB)	73.07(±0.45)	68.69(±0.56)
+RGCN	72.69(±0.19)	68.41(±0.66)
+GconAttn	72.61(±0.39)	68.59(±0.96)
+KagNet	73.47(±0.22)	69.01(±0.76)
+RN	74.57(±0.91)	69.08(±0.21)
+MHGRN	74.45(±0.10)	71.11(±0.81)
+QA-GNN	76.54(±0.21)	73.41(±0.92)
+CPE(ours)	78.92(±0.17)	75.01(±0.32)

Table 4: Performance comparison on Commonsense QA official split

Methods	Development	Test
RoBERTa	78.4	72.1
RoBERTa-IR	78.9	72.1
XLNet+Graph Reasoning	79.3	75.3
ALBERT+Path Generator	78.42	75.6
PEAR	78.42	76.1
RoBERTa+QAGNN	-	76.1
RoBERTa+MHGRN	-	76.5
ALBERT	80.5	73.5
T5	-	78.1
UnifiedQA	-	79.1
ALBERT+KCR	-	79.5
ALBERT+DEKCOR	84.7	80.7
ALBERT+CPE(ours)	84.93	80.13

Table 5: Accuracy on the test set of OpenBookQA

Methods	Test
RoBERTa-large(w/o KB)	64.80(±2.37)
+RGCN	62.45(±1.57)
+GconAttn	64.75(±1.48)
+RN	65.20(±1.18)
+MHGRN	66.85(±1.19)
+QAGNN	70.58(±1.42)
+CPE	71.00(±0.91)

3.5 Ablation Experiments

To further evaluate our method, we perform ablation studies on CommonsenseQA's official development set. The paraphrase association graph, entity paraphrases, and context are removed to further evaluate the contributions of the various modules of the CPE model.

[3] https://github.com/jessionlin/csqa/

Table 6: Ablation results on the development set of CommonsenseQA

Methods	Development
CPE	84.93
-PAG	81.65
-PA	83.98
-Triple-text	80.75
-KBs	79.52

(1)-PAG: The paraphrase association graph is removed;

(2)-PA: The WordNet and Wiktionary are removed, and only ConceptNet is used;

(3)-Triple-text: The ConceptNet removed, and only WordNet and Wiktionary are used;

(4)-KBs: All KBs are removed, and only fine-tune on ALBERT.

The experimental results of Table 6 show that the accuracy dropped by 3.28% after the removal the paraphrase association graph, indicating that the entity paraphrase will have some effect on the non-associated entity and proving the validity of the paraphrase association graph. Removing the WordNet and Wiktionary, only the triple-text extracted from ConceptNet is used, and the accuracy decreased by 0.95%, proving that entity paraphrase helps the model to understand triple-text effectively; The accuracy dropped by 5.41% after removing all external knowledge, proving that the external knowledge bases provides some useful clues to commonsense question to help the MRC model improve the accuracy of answers.

As shown in table 7, we further analyze the experimental results of different models with two examples. For the first example, without introducing any external knowledge base, the model (-KBs) gets the wrong answer "easter"; introducing ConceptNet, only choice E "give gift" extracts most relevant triple-text on the ConceptNet, and the model (-PA) also gets the wrong answer "give gift"; introducing WordNet and Wiktionary, although both the paraphrase of "chrismas" and choice A "halloween" mention a specific time, the two models (-PAG and -Triple-text) don't get the correct answer due to the lack of the paraphrase association graph; Finally, we further incorporate the paraphrase association graph into the model. The Association graph masks the relationship between context-paraphrase and non-associated entities, preventing changing the meaning of the other non-associated entities. As a result, the model (CPE) gets the correct answer "halloween".

Table 7: Example analysis, each example gives triple-text and context-paraphrase extracted from KBs

Question:
If it is **Chrismas** time what came most recently before?
Choices:
A.**halloween** B.summer C.easter
D.kwanza E.give gift

Triple-Text:
E:christmas would make you want to give gift.

Paraphrase:
chrismas: period extending from Dec. 24 to Jan. 6.
halloween: The eve of all Hallows' Day; October 31st. ...
summer: the warmest season of the year.
easter: a Christian celebration of the Resurrection of Christ. ...
gift: something acquired without compensation.

Prediction:
CPE: halloween **-PAG:** easter **-PA:** give gift
-Triple-text: easter **-KBs:** easter

Question:
What type of non-vegetarian soup is one likely to find a **potato**?
Choices:
A.beef stew B.own kitchen C.**clam chowder**
D.kitchen cabinet E.pantry

Triple-text:
A: you are likely to find potato in beef stew.
B: you are likely to find potato in own kitchen.
C: you are likely to find potato in clam chowder.
D: you are likely to find potato in kitchen cabinet.
E: you are likely to find potato in pantry.

Paraphrase:
potato: an edible tuber native to South America. ...
beef: cattle that are reared for their meat.
stew: food prepared by stewing especially meat or fish with vegetables.
kitchen: a room equipped for preparing meals.
clam chowder: A type of chowder made from clams and usually potatoes.
kitchen cabinet: Built-in cabinet found in a kitchen.
pantry: a small storeroom for storing foods or wines.

Prediction:
CPE: clam chowder **-PAG:** beef stew **-PA:** beef stew
-Triple-text: beef stew **-KBs:** beef stew

In the second example, the question mentions "potato". The paraphrase of "clam chowder" also shows that it is type of chowder made from clams and usually potatoes. However, when the interpretation of the association graph is not considered, the model pays more attention to the relationship between "non-vegetarian" and "beef", resulting in all models except model (CPE) choosing the wrong answer "beef stew".

4 Conclusions

In this paper, we propose a context-paraphrase enhanced commonsense question answering method for the CQA task. First, we extract triples and entity paraphrases in KBs based on question and choice entities, the triples is converted to triple-texts based on a relational template. Then, we construct a HTP graph based on the relationships between the triple-texts, entities, and the entity paraphrases, and retrieve the most relevant triple-text based on the HTP graph. Next, we construct a paraphrase association graph and incorporate it into the answering process of the MRC model. Finally, we verify our method on CommonsenseQA and OpenBookQA, and further prove through ablation experiments that entity paraphrase is effective for improving CQA tasks. However, an entity generally has more than one paraphrase contained in WordNet, Wiktionary, and other knowledge bases, and the paraphrase of non-conforming scenarios will affect the judgment of the MRC model. Therefore, the next focus is filtering out the entity paraphrase that meets the current question scenario from many paraphrases.

References

Yonatan Bisk, Rowan Zellers, Ronan LeBras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: reasoning about physical commonsense in natural language. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence*, pages 7432–7439.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yanlin Feng, Xinyue Chen, Bill Yuchen Lin, Peifeng Wang, Jun Yan, and Xiang Ren. 2020. Scalable multi-hop relational reasoning for knowledge-aware question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1295–1309, Online. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Bill Yuchen Lin, Xinyue Chen, Jamin Chen, and Xiang Ren. 2019. KagNet: Knowledge-aware graph networks for commonsense reasoning. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2829–2839, Hong Kong, China. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Shangwen Lv, Daya Guo, Jingjing Xu, Duyu Tang, Nan Duan, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, and Songlin Hu. 2020. Graph-based reasoning over heterogeneous external knowledge for commonsense question answering. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8449–8456. AAAI Press.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Lin Qiu, Yunxuan Xiao, Yanru Qu, Hao Zhou, Lei Li, Weinan Zhang, and Yong Yu. 2019. Dynamically fused graph network for multi-hop reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6140–6150, Florence, Italy. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. 2017. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the*

9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European semantic web conference*, pages 593–607. Springer.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *International Conference on Learning Representations*.

Peifeng Wang, Nanyun Peng, Filip Ilievski, Pedro Szekely, and Xiang Ren. 2020. Connecting the dots: A knowledgeable path generator for commonsense question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4129–4140, Online. Association for Computational Linguistics.

Xiaoyan Wang, Pavan Kapanipathi, Ryan Musa, Mo Yu, Kartik Talamadupula, Ibrahim Abdelaziz, Maria Chang, Achille Fokoue, Bassem Makni, Nicholas Mattei, and Michael Witbrock. 2019. Improving natural language inference using external knowledge in the science questions domain. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7208–7215. AAAI Press.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yichong Xu, Chenguang Zhu, Ruochen Xu, Yang Liu, Michael Zeng, and Xuedong Huang. 2021. Fusing context into knowledge graph for commonsense reasoning. In *The 59th Annual Meeting of the Association for Computational Linguistics (ACL)*.

Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. 2021. QA-GNN: Reasoning with language models and knowledge graphs for question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 535–546, Online. Association for Computational Linguistics.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3363–3369, Hong Kong, China. Association for Computational Linguistics.